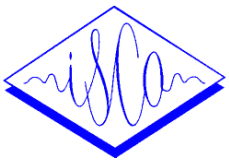

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

INTELLIGIBILITY OF AUDIO-VISUALLY DESYNCHRONISED SPEECH: ASYMMETRICAL EFFECT OF PHONEME POSITION

P.M.T. Smeele[1], A.C. Sittig[1] and V.J. van Heuven[2]

[1] Dept. of Industrial Design Engineering, Delft University of Technology,
Jaffalaan 9, 2628 BX Delft, The Netherlands, Fax: 31-15-787316
[2] Phonetics Laboratory, Dept. of Linguistics, Leyden University,
P.O. Box 9515, 2300 RA Leyden, The Netherlands

ABSTRACT

In an earlier experiment we studied the effect of desynchronisation of visual and auditory information on the perception of nonsense CVC-words. Predictably, percent correctly recognized words was generally smaller in asynchronous conditions than in the synchronous condition. However, an asymmetrical effect was found: subjects' performance was poorer in the condition where audition preceded vision than when vision preceded audition. In order to gain insight into this asymmetrical effect and its causes, we have now examined subjects' responses to nonsense words on the phoneme level. Analysis of the individual phonemes which made up the original CVC-stimuli indicates that identification of the initial consonant was worse when vision preceded audition than vice versa. In contrast to this, identification of both the vowel and the final consonant was poorer when audition preceded vision. This effect of phoneme position may serve to explain why perception deteriorates more when audition precedes vision than when vision leads.

10.21437/ICSLP.1992-19

I. INTRODUCTION

Our research aims towards a coherent account of human bimodal (audio/visual) speech perception. We are specifically interested in the integration process of auditory and visual information.
Earlier research has shown that visual information provided by a speaker's face and movements of his mouth can improve the intelligibility of speech, relative to conditions where only auditory information is presented. This improvement is most easily demonstrated in situations where the auditory signal is degraded, e.g. due to a hearing impairment or to the presence of noise [9,15].
Whether the speech materials presented are sentences [10,15], lexical words [4,14], or nonsense words [2,12], intelligibility is superior in the presence of additional visual information. The results of studies with nonsense words are of particular importance, since they indicate that a contribution of vision exists regardless of contextual information or of the lexical status of the stimuli.
To gain more insight into the audiovisual integration process, experiments were performed in which conflicting auditory and visual information were presented [6,7,8]. These studies show that vision strongly influences perception, in certain cases even leading to percepts that were presented neither auditorily nor visually. It was first demonstrated by McGurk and MacDonald [8] that observers, being presented with an auditory [ba] and a visual [ga], perceive [da]. This has since been known as the "McGurk effect". Obviously, observers are able to integrate conflicting visual and auditory information. We intend to explore further the conditions and limitations of this integration process.
Another way of studying the integration process is using the experimental method of manipulating the synchronisation between the audio and visual channels. Campbell and Dodd used desynchronisations of 400, 800 and 1600 ms where vision always preceded audition [4]. They found that phoneme identification was invariably better in the asynchronous conditions than in the auditory-only control condition. This indicates that visual information can still be successfully integrated with audition even when there is a severe time delay.
Dixon and Spitz studied the detection of desynchronisation between audition and vision, rather than its effect on intelligibility [5]. In their experiment subjects watched and heard a human speaker, and watched and heard a nonlinguistic (and acoustically transient) event such as a hammer hitting a peg. They found that asynchrony was detected sooner when sound preceded visual information than vice versa. Furthermore, there was a greater sensitivity to desynchronisation in the 'hammer' condition than in the 'speech' condition. Detection thresholds were 75 ms (sound first) and 188 ms (image first) for the 'hammer' condition versus 131 ms (sound first) and 258 ms (image first) for the 'speech' condition.
We ourselves studied the effect of desynchronisation on the intelligibility of isolated Dutch nonsense CVC-words [13]. Desynchronisations ranged from -280 ms (audition before vision) to +280 ms (vision before audition). Figure 1 summarizes the results of this experiment.

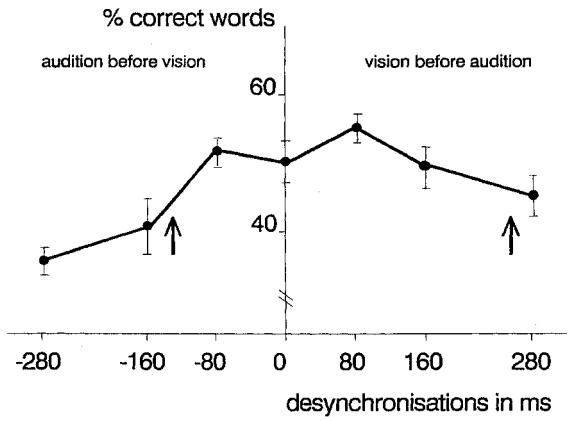

Figure 1. Percent correctly reported nonsense words for 7 desynchronisation conditions. Scores are averaged over 10 subjects, whiskers indicate standard error. Arrows indicate the speech asynchrony detection thresholds found by Dixon and Spitz [5].

We observe that intelligibility is generally poorer for conditions with larger asynchronies. Moreover, asymmetry is apparent in our results: intelligibility suffers more from negative desynchronisation (audition before vision) than from positive desynchronisation (vision before audition). Clearly then, the asymmetry in our intelligibility study is in line with the results obtained in the asynchrony detection paradigm [5]. Dixon and Spitz also found an asymmetrical performance in the detection tasks. Although detection tasks are not equivalent to intelligibility tasks, this parallel indicates, to us, that for both tasks the audiovisual integration mechanism responsible for the asymmetry is the same.
To study the asymmetrical effect of desynchronisation more closely, we have now analysed the subjects' responses to the individual phonemes that made up the CVC-stimuli.
Before we proceed this analysis we shall briefly recapitulate the experimental setup.

II. METHOD

2.1 Experimental Setup

Ten normally hearing subjects, native Dutch speakers, were presented with video recordings of a face of a female (native Dutch) speaker pronouncing nonsense words. These nonsense words were isolated, phonetically balanced Dutch CVC-syllables. The nonsense words were presented at a loudness level of 55-60 dB(A) against background speech noise of 60-65 dB(A). Desynchronisation conditions were: -280, -160, -80 ms (audition before vision), 0 ms (audition and vision synchronously), +80, +160, +280 ms (vision before audition). In each condition subjects were presented with a set of 50 nonsense stimuli. A different set was used for each condition. Subjects sat at a distance of 1.2 m facing a video display monitor (52 cm). They were instructed to carefully watch and listen to the speaker, and to write down in standard Dutch orthography, for each item, the sounds they thought the speaker had said.

2.2 Phoneme Analysis

We analysed the subjects' responses to the individual phonemes that made up the nonsense CVC-words. For each subject we calculated the percentage of correctly identified phonemes within each of the 7 desynchronisation conditions: -280, -160, -80, 0, +80, +160, +280 ms.
The phoneme scores were partitioned symmetrically into three groups:

i) the two extreme negative desynchronisation conditions (-280 and -160 ms) together,
ii) the three most synchronous conditions (-80, 0, +80 ms), and
iii) the two extreme positive desynchronisation conditions (+160 and +280 ms) together.

This partitioning was done because the behaviour of all the subjects showed this grouping. Within the three defined groups there was no clear trend for individual subjects. The three groups were significantly different, $F(\text{group i vs. ii}) = 32.05$, $p<<0.001$; $F(\text{group i vs. iii}) = 9.87$, $p<0.005$ (reflecting the asymmetry); $F(\text{group ii vs. iii}) = 4.93$, $p<0.05$.

III. RESULTS

We calculated percent correct identification of the initial consonant, vowel, and final consonant of CVC-words that were presented with asynchronous sound and vision. The results are shown in figure 2.

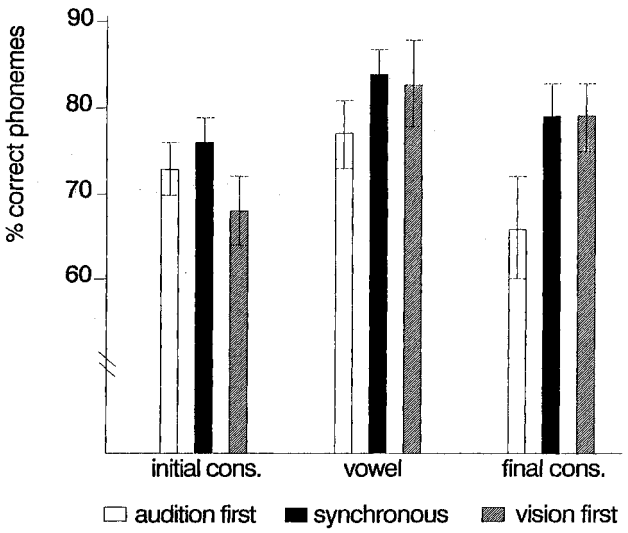

Figure 2. Percent correctly identified phonemes for three types of audiovisual asynchrony conditions. Scores are averaged over 10 subjects, whiskers indicate standard deviations.

For all syllable positions (initial consonant, vowel, final consonant) performance was best in the synchronous condition.
A comparison between the positive and the negative desynchronisation conditions reveals that identification of the initial consonant is significantly poorer when vision precedes audition than when audition precedes vision ($p<0.025$). In contrast to this, both the vowel and the final consonant are identified less successfully when audition precedes vision than when vision leads ($p<0.005$ and $p<0.001$ respectively).
It appears that the position of a phoneme in a word influences the extent to which desynchronisation affects intelligibility. Since there are, in the negative desynchronisation condition, two phonemes (vowel and final consonant) that are identified more poorly and only one (initial consonant) that is identified better, the intelligibility of the complete word will-ceteris paribus - be better when vision leads than when audition leads. Indeed, the probability P of correctly reporting the CVC-word, given by the product of the probabilities of correctly identifying the separate phonemes, yields $P = 0.73 * 0.77 * 0.66 = 0.37$ for the audition-before-vision condition, and $P = 0.68 * 0.83 * 0.79 = 0.45$ for the vision-before-audition condition.

IV. DISCUSSION

Presenting audition and vision out-of-synchrony generally affects the intelligibility of nonsense words (Fig. 1; [13]). However, there is an asymmetrical range of desynchronisations where the intelligibility is not impaired: desynchronisations in the range from roughly -100 ms (audition before vision) to about +200 ms (vision before audition) fail to have a significant effect. This range is very similar to the range of asynchronies between auditory and visual speech that remained undetected in the experiments by Dixon and Spitz [5]. Their speech asynchrony detection thresholds are indicated by arrows in figure 1. We suggest that small desynchronisations do not affect word intelligibility because they cannot be perceived by the observers.
For the larger desynchronisations presented we also found an asymmetrical effect; the decrease of percent correctly reported words was larger for negative desynchronisations (audition before vision) than for physically equivalent positive desynchronisations (vision before audition). This can also be thought to reflect the asymmetrical effect of asynchrony on the detectibility. The asymmetry in intelligibility we found, would then originate from differences between the processing of auditory and visual information, either in the early stages of perception or in short-term information storage. Such an asymmetry can also be functionally meaningful: in the real world sound often (strictly speaking: always) reaches the observer's ear later than the visual information related to the same event reaches the eye, but the reverse will never happen.
Is the finding that desynchronisation of vision and audition is sooner detected for a hammer hitting a peg than for speech [5] conflicting with this idea? We think not. Firstly, it is significant that the detection thresholds are higher by approximately the same absolute amount (120 ms) for the positive desynchronisation as compared to the negative desynchronisation in both cases. This again suggests that early perceptual processes cause the differences in thresholds. Secondly, the difference in absolute detection thresholds for speech and the 'hammer' condition can readily be explained by a more exact definition in time for the latter event.
The difference in detection thresholds for asynchrony can thus account for the asymmetry in the intelligibility task on the word level. However, this difference does not explain the present analysis of the responses at the phoneme level, viz., that the identification of the initial consonant, in contrast to the vowel and the final consonant, was poorer in the condition where vision preceded audition than vice versa.
We would like to understand why the identification of the initial consonant is affected by asynchrony in a different way than are the vowel and the final consonant. We distinguish two different approaches to this problem. The first one focusses on the role of the position of the phoneme within the syllable. In the synchronous condition the vowel is the best identified phoneme, followed by the final consonant and, finally, the initial consonant. This order of phoneme intelligibility is what we usually find in identification tasks involving CVC monosyllables in Dutch [1,3,11]. These results may partly be explained from acoustic differences between vowels and consonants: vowels generally exceed background noise by a greater margin than consonants. Also, there the phoneme inventories differ for vowels: uncertainty is largest for initial consonants (19 possibilities), intermediate for vowels (15 stressable vowels), and least for final consonants (13 possibilities). We observe the same order of phoneme intelligibility in the condition where vision precedes audition. However, in the audition-before-vision condition we see a different order: the initial consonant can here be identified better than can the final consonant. We would like to be able to compare the performance in this condition with the performance in the auditory-only condition. From this comparison we could learn if vision contributes at all in the conditions with the poorest phoneme identification. These experiments have yet to be run.
The second type of approach takes into account the importance of the temporal coincidence of the visual and auditory signals. For instance, our observations are congruent with the following tentative explanation: when vision precedes audition, the initial part of the visual signal is not interpreted as speech information (because of the lack of simultaneous auditory speech information), and, therefore, does not contribute to intelligibility.
New experiments, in which continuous speech will be presented, rather than isolated words, will clarify if temporal coincidence of vision and audition or position of the phoneme within a syllable are the relevant parameters to describe the process of

audiovisual integration in speech perception.

REFERENCES

[1] Bezooijen, R. van, & Pols, L.C.W. (1992) "Evaluation of text-to-soeech conversion for Dutch", in: V.J. van Heuven, L.C.W. Pols (Eds.): Analysis and synthesis of speech, towards high-quality text-to-speech generation, Berlin: Mouton de Gruyter (in press).

[2] Binnie, C.A., Montgomery, A.A., & Jackson, P.L. (1974) "Auditory and visual contributions to the perception of selected English consonants for normally hearing and hearing-impaired listeners", in: H. Birk Nielsen & E. Kampp (Eds.), Visual and audio-visual perception of speech (Scandinavian Audiology Supplement 4, 181-209). Stockholm: Almquist & Wiksell

[3] Boeschoten, J.A. van (1989) "Intelligibility of sounds in Dutch spoken by Turks", Doctoral dissertation, Leyden University

[4] Campbell, R., & Dodd, B. (1980) "Hearing by eye", Quarterly Journal of Experimental Psychology 32, 85-99

[5] Dixon, N.F., & Spitz, L (1980) "The detection of auditory visual desynchrony", Perception 9, 719-721

[6] Green, K.P., & Kuhl, P.K. (1989) "The role of visual information in the processing of place and manner features in speech perception", Perception & Psychophysics 45, 34-42

[7] Massaro, D.W., & Cohen, M.M. (1983) "Evaluation and integration of visual and auditory information in speech perception", Journal of Experimental Psychology: Human Perception & Performance 9, 753-771

[8] McGurk, H., & MacDonald, J. (1976) "Hearing lips and seeing voices", Nature 264, 746-748

[9] Miller, G.A., & Nicely, P.E. (1955) "An analysis of perceptual confusions among some English consonants", Journal of the Acoustical Society of America 27, 338-352

[10] Reisberg, D., McLean, J., & Goldfield, A. (1987) "Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli", in: B. Dodd & R. Campbell (Eds.), Hearing by eye: The psychology of lip-reading, 97-113. London: Erlbaum

[11] Ringeling, J.C.T. (1984) "Reducing redundancy in in normal, soft and whispered speech: a study on native and near-native perception, Doctoral dissertation Utrecht University

[12] Smeele, P.M.T., & Sittig, A.C. (1991) "The contribution of vision to speech perception", Proceedings of 2nd European Conference on Speech Communication and Technology, Eurospeech 91, Genova, 1495-1497

[13] Smeele, P.M.T., & Sittig, A.C. (1991) "Effects of desynchronization of vision and speech on the perception of speech: preliminary results", CCITT Brazil Conference sept.'91, Stgrp. XII Wp. XII/2 and XII/3, Contribution D.81

[14] Sumby, W.H., & Pollack, I. (1954) "Visual contribution to speech intelligibility in noise", Journal of the Acoustical Society of America 26, 212-215

[15] Summerfield, Q. (1979) "Use of visual information for phonetic perception", Phonetica 36, 314-331